효율적인 진화알고리즘을 이용한 적응형 퍼지 분류 규칙 생성*

류정우°, 김성은**, 김명원**
한국전자통신연구원 지능형로봇연구단 지식및추론연구팀°
숭실대학교 컴퓨터학부**
ryu0914@etri.re.kr°, babystep@ssu.ac.kr, mkim@comp.ssu.ac.kr

Generating Adaptive Fuzzy Classification Rules using An Efficient Evolutionary Algorithm

Joung Woo Ryu°, Sung Eun Kim**, Myung Won Kim**
Intelligent Robot Research Division, Knowledge & Inference Research Team, ETRI°
Department of Computer Science, Soongsil University**

요 약

데이터 특성이 연속적이고 애매할 때 퍼지규칙으로 분류 규칙을 표현하는 것은 매우 유용하고 효과적이다. 그러나 일반적으로 정확하지 않은 데이터 특성에 대해서 소속함수를 결정한다는 것은 어려운 일이다. 본 논문에서는 진화알고리즘을 이용하여 효과적인 퍼지 분류 규칙을 자동으로 생성하는 방법을 제안한다. 제안한 방법에서 규칙의 정확성과 이해성을 고려하여 최적화된 소속함수를 생성하기 위해 진화알고리즘을 사용한다. 먼저 지도 군집화로 진화를 위한 초기 소속함수를 생성한다. 진화알고리즘은 전역적 최적 해를 찾는데 효과적이다. 그러나 시간에 대한 효율성이 낮다. 특히 모델 최적화 문제에서는 개체 평가 단계에서 많은 시간이 소요된다. 따라서 본 논문에서는 전체 데이터를 여러 개의 부분 데이터들로 나누고 개체들은 전체 데이터 대신 매번 부분 데이터를 임의적으로 선택하여 개체를 평가함으로써 수행 시간을 단축시킬 수 있는 진화 방법을 제안한다. 제안한 퍼지 분류 규칙 생성 방법의 타당성을 검증하기 위한 실험 데이터로 UCI에서 제공하는 데이터들을 사용하였으며, 실험 결과는 기존 방법에 비해 평균적으로 더 효과적임을 확인하였다.

1. 서론

분석할 데이터가 연속적인 특성을 갖는 수치형 속성일 경우, 퍼지 분류 규칙이 일반 분류 규칙보다 정확성과 이해성이 높은 분류 규칙들을 생성할 수 있다.

수치형 속성에서 일반 분류 규칙을 생성하기 위해 기호형 속성으로 변화 시켜주는 이산화 (discretization) 전처리 과정을 수행해야만 한다. 이산화는 수치형 속성의 범위를 여러 구간으로 나누고 각 구간에 레이블 (label)을 정의함으로써 기호형 속성으로 변화시켜 주는 전처리 방법이다. 이산화 될 때, 구간들 간의 경계가 명확하게 정의됨으로써 경계 부근에서 미세한 값의 변화에도 결과가 크게 달라지는 문제점 (sharp boundaries problem)이 있다. 그러나 퍼지 분류 규칙을 생성하기 위해 수치형 속성을 구간으로 구분할 경우, 구간에 포함되는 데이터들이 [0,1]의 값을 가질 수 있도록 소속함수를 정의하고 경계 부근에 존재하는 데이터들이 동시에 근접한 소속함수에 포함될 수 있도록 소속함수들을 중첩시킴으로써 경계 부분의 애매성을 처리한다. 따라서 퍼지 분류 규칙의 정확성과 이해성은 소속함수의 변수와 개수에 따라 달라진다. 이와 같이 문제에 따라 최적의 소속함수의 변수와 개수를 찾는 것은 NP-complete 문제이다. [1]에서는 최적의 퍼지 의사결정 트리를 생성하는 것이 NP-complete 문제라는 것을 증명하였다.

본 논문에서는 문제의 특성에 따라 적응형 퍼지 분류 규칙을 생성할 수 있는 방법을 제안한다. 분류 문제에서 문제의 특성은 클래스 분포로 나타낼 수 있다. 그러므로 클래스 분포에 따라 공간을 분할하고 분할된 공간을 이용하여 초기 소속함수를 생성한다. 생성된 초기 소속함수는 퍼지 의사결정 트리에 의해 퍼지 분류 규칙을 생성하고 생성된 퍼지 분류 규칙의 정확성과 이해성을 향상시키기 위해 진화알고리즘으로 초기 소속함수를 진화시키는 방법을 제안한다. 또한, 진화알고리즘의 시간에 대한 효율성을 높이기 위해 개체 평가 단계에서 전체 데이터를 이용하는 대신 전체 데이터를 여러 개의 부분 데이터들로 나누고, 개체를 평가할 때마다 부분 데이터들 중 임의로 한 부분 데이터를 선택하여 평가함으로써 수행 시간을 단축시킬 수 있는 진화 방법을 제안한다.

2. 관련연구

기존 퍼지 분류 규칙을 생성하는 방법은 크게 두 가지로 구분할 수 있다. 1) 사전에 각 속성 별로 소속함수의 변수와 개수를 정의한 다음 정의된 소속함수에 의해 생성될 수 있는 모든 규칙을 생성하고, 이들 중 적합한 퍼지 분류 규칙들을 찾아 생성하는 방법[2,3,4]과 2) 사전에 퍼지 분류 규칙의 형태와 개수를 정의하고 정의된 퍼지 분류 규칙으로 정확성을 높일 수 있도록 소속함수를 학습함으로써 생성하는 방법[5,6]이다. 이러한 방법들은 소속함수 개수를 고정시키거나 혹은 규칙의 개수를 고정시킨 상태에서 퍼지 분류 규칙을 생성함으로 문제의 특성에 따라 성능의 차이가 크게 나타난다.

[2]에서는 소속함수의 변수와 개수를 사전에 정의하고 정의된 소속함수를 이용하여 문제 영역을 그리드 분할한 후, SLAVE 평가 기준을 적용하여 후보 규칙들을 추출한다. 추출된 후보 규칙들 중에서 유용한 규칙들을 선택할 수 있는 유전자 규칙 선택 방법을 제안하였고, [3]에서는 사전에 직관적으로 정의된 삼각형 소속함수를 학습시키는 방법으로 오류 보정 학습 방법을 제안하였다. [4]에서는 사전에 대칭형 삼각형 소속함수를 정의하고, 이를 통해 공간을 그리드 분할한 후, 데이터가 포함된 격자만을 규칙으로 생성한다. 생성된 규칙들은 규칙의 확신도 (certainty factor)만을 학습시킴으로써 퍼지 분류 규칙을 생성한다.

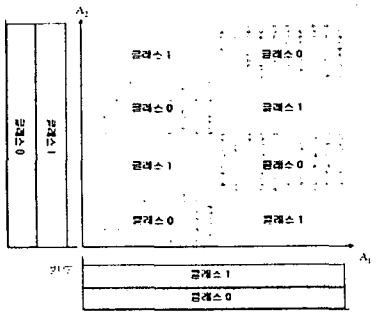
[5]에서는 초기 퍼지 분류 규칙을 생성하기 위해 퍼지 군집화로 공간을 분할하고 각 데이터들이 군집에 포함되는 소속 정도를 이용하여 초기 소속함수를 생성한다. 생성되는 초기 퍼지 분류 규칙은 모든 속성이 조건부의 조건으로 이루어진 완전 퍼지 분류 규칙 (complete fuzzy classification rule)이다. 따라

* 본 연구는 한국학술진흥재단 선도연구자지원사업에 의해 수행되었습니다. (과제번호: 2004-041-D00627)

서 조건부를 간소화하기 위해 퍼지 분류 규칙 간소화 단계를 제안하였고, 진화알고리즘으로 소속함수를 조율시킴으로써 생성되는 퍼지 분류 규칙의 정확성을 향상시켰다. [5]에서 제안한 방법은 초기 퍼지 분류 규칙 생성 단계에서 군집의 개수를 클래스 개수로 설정하기 때문에 생성될 수 있는 최대 규칙의 개수는 클래스 개수가 된다. 따라서 <그림 1>과 같이 한 클래스가 두 개 이상의 군집으로 분포 되었을 경우에는 정확한 규칙을 생성할 수 없다. [6]에서는 퍼지 분류 규칙 생성 방법이 [5]와 같다. 단, 군집화 대신 C4.5를 이용하여 공간을 분할하여 초기 소속함수를 생성하였다는 점에서 다르다.

지금까지 서술한 방법이외에 퍼지 분류 규칙 생성 방법으로는 의사결정 트리에 퍼지 이론을 적용한 퍼지 의사결정 트리 방법이 있다. 퍼지 의사결정 트리 생성 방법 중 Janikow에 의해 제안된 FID (Fuzzy ID3)[7]은 수치형 속성에 있어 사전에 소속함수 정의 여부에 상관없이 모두 처리 가능하다. 특히, 사전에 소속함수가 정의되지 않은 속성을 잠재적 속성 (potential attribute)이라 정의하고 있다. 잠재적 속성을 처리 할 때, C4.5에서와 같이 무질서도 (entropy)가 가장 낮은 분할 점 (split point)을 찾고 이를 기준으로 두 개의 사다리꼴 소속함수를 생성한다. 또한 트리를 생성할 때 한 경로에서 동일한 잠재적 속성을 반복해서 선택하여 자식 노드를 생성할 수 있다. 따라서 잠재적 속성에 많은 소속함수가 생성될 수 있기 때문에 최대 생성될 수 있는 소속함수 개수를 사전에 결정해야만 한다.

[6][7]과 같이 퍼지 분류 규칙을 생성할 때 속성 별로 무질서도를 이용할 경우 <그림 1>과 같이 클래스별 균등 분포일 경우, 즉 클래스들이 모든 영역에서 클래스별로 균등하게 분포되어 있을 경우에는 규칙을 생성할 수 없다.



<그림 1> 클래스별 균등 분포

3. 효율적인 진화알고리즘을 이용한 적응형 퍼지 분류 규칙 생성 방법

본 논문은 적응형 퍼지 분류 규칙을 생성하기 위해 데이터의 클래스 분포에 따라 초기 소속함수의 변수와 개수를 결정하고 진화알고리즘을 이용하여 생성되는 퍼지 분류 규칙들의 이해성과 정확성이 높도록 소속함수를 진화시킨다. 소속함수는 가장 보편적으로 사용하는 삼각형 소속함수로 정의한다.

이와 같이 모델을 최적화하는데 있어 진화알고리즘을 사용하게 되면, 개체 평가 단계에서 수행 시간이 많이 소요됨으로써 시간에 대한 효율성이 떨어진다. 따라서 제안한 방법에서는 전체 데이터를 이용하여 개체를 평가하는 대신 사전에 분할한 부분 데이터들로부터 평가할 때 마다 임의로 하나를 선택하여 개체를 평가함으로써 진화알고리즘의 효율성을 향상시킨다.

제안한 방법은[8]에서 제안한 지도 군집화 방법을 이용하여 클래스 분포에 따라 최적화된 군집의 개수를 찾고, 생성된 군집들의 중심 좌표들을 각 속성 별로 투영하여 초기 삼각형 소속함수의 꼭지점으로 설정한다. 이 때 중심과 중심사이에 데이터가 없거나 데이터의 클래스가 동일한 경우, 두 중심을 중심간의 중앙값으로 대치함으로써 속성 별로 불필요한 소속함수를 생성하지 않도록 한다.

다음은 효율적인 진화알고리즘으로 초기 소속함수를 최적화시키는 단계에 대해서 서술한다.

3.1 초기 개체집단

개체는 속성들의 속성 값에 해당하는 퍼지집합의 소속함수 집합을 나타내고, 속성마다 정의된 소속함수의 개수가 다르기 때문에 개체를 가변길이 실수 표현으로 인코딩한다.

초기 개체집단의 소속함수들은 앞에서 기술한 것처럼 초기 삼각형 소속함수의 꼭지점들을 이용하여, 다음과 같은 두 가지 제약조건에 따라 소속함수의 양 끝점을 임의적으로 생성함으로써 생성된다.

1) 분할 조건: 이웃하는 소속함수들은 반드시 중첩되게 한다.
2) 탐색 공간 조건: 현재 상태에서 소속함수의 좌표점이 변경될 수 있는 범위를 정의한다.

3.2 개체 평가

소속함수 최적화를 위한 평가 단계는 분류 모델 생성 과정과 평가 과정으로 나누어진다. 분류 모델 생성 과정은 개체에 인코딩된 소속함수들을 이용하여 [9]에서 기술한 퍼지 의사결정 트리에 의해 퍼지 분류 규칙을 생성함으로써 이루어진다. 모델 평가 과정은 mamdani의 Min-Max 추론 방법을 통해 평가된다.

이와 같이 개체에 의해 생성된 분류 모델을 생성하고 평가함으로써 개체에서 표현하고 있는 소속함수들에 의해 정확성과 이해성이 높은 퍼지 분류 규칙을 생성할 수 있는지 평가하기 위한 적합도 함수를 <식 1>와 같이 정의한다.

$$Fit_i = wA_i(\tau) + (1-w)(1-C_i(\tau)),\ 0 \le w \le 1 \qquad \text{<식 1>}$$

여기서 τ는 퍼지 분류 규칙으로 구성된 분류 모델이다. 따라서 $A_i(\tau)$는 i개체에 의해 생성된 분류 모델의 정확성을 확인하기 위한 인식률 (accuracy)을 나타내고, $C_i(\tau)$는 이해성을 확인하기 위한 복잡성 (complexity)을 나타낸다. 복잡성은 <식 2>과 같이 규칙의 개수로 정의한다.

$$C_i(\tau) = \begin{cases} \dfrac{R_i(\tau)}{|N|} & , R_i(\tau) \le |N| \\ 0 & , R_i(\tau) > |N| \end{cases} \qquad \text{<식 2>}$$

여기서 N은 전체 데이터 집합이고, $R_i(\tau)$은 개체 i에 의해 생성된 분류 모델 τ에 포함된 규칙의 개수를 나타낸다. 본 논문에서는 분류 모델이 가질 수 있는 최대 규칙의 개수를 전체 데이터 개수로 설정한다. 만약 분류 모델을 구성하고 있는 규칙의 개수가 전체 데이터 개수보다 크다면 모델은 필요 이상 규칙을 가지고 있음으로 모델의 복잡성을 0으로 계산한다.

3.3 진화 연산

가변길이 실수 표현으로 인코딩한 개체에 대해 진화 연산은 가우시안 돌연변이 연산과 전체 산술 교배 (whole arithmetic crossover)[5], 휴리스틱 교배 (heuristic crossover)[5], 속성 교체 교배 (attribute change crossover)로 정의한다.

교배 연산은 속성 간의 소속함수 개수가 같은 경우와 같지 않은 경우로 나누어 적용한다. 전자의 경우에는 전체 산술 교배와 휴리스틱 교배를 임의로 선택하여 적용한다. 후자의 경우에는 속성 교체 교배를 적용한다.

4. 실험

제안한 퍼지 분류 규칙 생성 방법의 타당성을 검증하기 위해 UCI에서 제공하는 벤치마크 데이터들을 이용하여 기존 방법과 비교하였다.

제안한 방법에서 퍼지 의사결정 트리를 이용하여 퍼지 분류

규칙을 생성하기 때문에 생성되는 규칙의 정확성과 이해성이 트리 생성에서 단말 노드 조건에 영향을 받게 된다. 따라서 <표 1>은 단말 노드 조건을 위한 임계값에 따라 생성되는 퍼지 분류 규칙의 정확성과 이해성을 보여준다. θ_d는 분할 공간에서 대표 클래스의 비율에 대한 임계값이고, θ_e는 분할 공간에 포함된 데이터의 소속정도 합의 비율에 대한 임계값을 의미한다. θ_d가 높을수록 θ_e가 낮을수록 생성되는 트리는 커질 수 있다.

<표 1> 단말노드 조건의 임계값에 따른 비교

데이터	w=0.1, θ_d=1.0, θ_e=0.02			w=0.1				
	인식률 (%)	규칙 개수	조건항 개수	인식률 (%)	규칙 개수	조건항 개수	θ_d	θ_e
iris	97.3	3	4	98.0	3	4	1.00	0.06
pima	75.0	2	2	75.6	3	4	0.85	0.01
bcw	97.0	2	4	96.6	2	4	1.00	0.01
wine	76.4	9	24	92.7	5	8	0.85	0.05
glass	78.9	54	114	69.6	26	48	0.85	0.04

<표 2>은 [10]에서 제안한 방법과 비교한 실험 결과이다. [10]에서는 정확하고 간결한 퍼지 의사결정 트리를 생성하기 위해 노드를 확장할 속성을 선택하는 과정에서 분류 애매성(classification ambiguity)이라는 척도를 제안하였고, 가중치 퍼지 규칙 (weighted fuzzy rules)을 이용한 추론 방법을 제안하였다. 그러나 사전에 각 속성 별로 세 개씩 직관적으로 삼각형 소속함수를 생성하였다.

<표 2> [10]과 제안한 방법과 성능 비교

데이터	인식률 (%)		단말노드 개수		노드 개수	
	[10]	제안방법	[10]	제안방법	[10]	제안방법
iris	97.0	97.3	9.7	3.0	16.0	4.0
rice	88.0	98.0	8.8	2.0	14.5	3.0
thyroid	86.0	94.8	7.8	3.0	15.2	5.0
pima	80.0	75.0	34.7	2.0	53.2	2.0

제안한 방법이 기존 방법 [10]보다 생성된 퍼지 분류 규칙이 더 정확하고 간결하다는 것을 알 수 있다. 따라서 직관적으로 생성한 소속함수보다 데이터 특성을 고려하여 생성한 소속함수가 보다 정확하고 간결한 퍼지 분류 규칙을 생성할 수 있다.

<표 3>은 [6]에서 제안한 방법과 비교한 실험 결과이다. [6]에서는 C4.5로 공간을 분할하여 초기 소속함수와 초기 퍼지 분류 규칙을 생성하고, 진화알고리즘으로 초기 규칙들을 진화시킴으로써 간소화된 정확한 퍼지 분류 규칙을 생성하였다.

<표 3> [6]과 제안한 방법과 성능 비교

데이터	인식률 (%)		규칙 개수		조건항의 개수	
	[6]	제안방법	[6]	제안방법	[6]	제안방법
iris	96.1	98.0	3.0	3.0	4.0	4.0
ionosphere	86.4	91.4	3.4	3.0	10.2	6.0
glass	66.0	69.6	19.2	26.0	90.8	48.0
pima	73.0	75.6	11.2	3.0	40.0	4.0
wine	91.2	92.7	3.6	5.0	8.8	8.0
bcw	96.8	96.6	2.0	2.0	4.0	4.0

일반적으로 제안한 방법에서 생성된 규칙들의 정확성과 이해성이 향상되었으나, glass 데이터 같은 경우 기존 방법보다 조건부가 짧은 규칙이 더 많이 생성되었다. 이는 제안한 방법에서 각 속성 별 초기 소속함수의 개수를 클래스 분포에 따라 결정하기 때문에 일반적으로 클래스 개수만큼 생성됨으로 발생되는 결과이다. 또한 [6]에서는 <그림 1>과 같은 데이터에 대해서는 규칙을 생성할 수 없다. 그러나 제안한 방법은 지도 군집화로 공간을 분할하여 초기 소속함수를 생성하기 때문에 규칙을 생성할 수 있다.

다음 실험은 본 논문에서 제안한 효율적인 진화알고리즘의 타당성을 확인하기 위해 두 가지 경우에 대해 실험하였다. 첫 번째 경우는 전체 데이터를 사전에 5개의 부분 데이터로 분할하였고, 나머지 경우는 10개의 부분 데이터로 분할하였다. 부분 데이터를 생성할 때, 전체 데이터에서 클래스 비율에 따라 비복원 추출로 생성하였다. <표 4>은 전체 데이터 개수가 많은 bcw와 pima 데이터에 대한 실험 결과이다. 부분 데이터 개수만큼 진화 시간이 단축되었으며, 그에 비해 생성된 퍼지 분류 규칙에 대한 정확도는 차이가 거의 없음을 알 수 있다. 표에서 분할 개수가 1이라는 것은 전체 데이터를 의미한다.

<표 4> 제안한 진화 방법의 효율성 실험 결과

데이터	분할 개수	분할된 평균 데이터 개수	개체별 평균시간 (초)	인식률 (%)	규칙 개수	조건항 개수	적합도
bcw	1	683.0	4.19	96.6	2.0	4.0	0.4967
	5	136.6	0.84	95.7	2.0	4.0	0.4962
	10	68.3	0.4	95.3	2.0	4.0	0.4965
pima	1	763.0	7.49	75.6	3.0	4.0	0.4858
	5	153.6	0.8	74.7	2.0	2.0	0.4861
	10	76.8	0.42	71.3	2.0	3.0	0.4845

5. 결론 및 향후 연구

여러 개의 벤치마크 데이터를 이용하여 기존 방법과 비교를 통해 제안한 방법이 보다 정확성과 이해성이 높은 퍼지 분류 규칙을 생성할 수 있다는 것을 확인하였고, 모델 최적화에 있어 진화알고리즘의 시간에 대한 효율성을 높일 수 있는 진화 방법을 제안하였다.

향후 연구로는 문제에 대한 도메인 정보를 활용한 휴리스틱 진화 연산을 연구하여 탐색 공간이 증가하여도 보다 빨리 최적의 해에 수렴될 수 있는 효율적인 진화알고리즘을 개발한다.

6. 참고문헌

[1] X. Z. Wang, B. Chen, G. Qian, F. Ye, "On the optimization of fuzzy decision trees," Fuzzy Sets Systems, Vol.112, No.2, pp.117-125, 2000

[2] Hisao Ishibuchi, Takashi Yamamoto, "Effects of Three-Objective Genetic Rule Selection on the Generalization Ability of Fuzzy Rule-Based Systems," Evolutionary Multi-Criterion Optimization, LNCS 2632, pp.608-622, 2003

[3] Nakashima, T. Nakai, G. Ishibuchi, H., "Improving the performance of fuzzy classification systems by membership function learning and feature selection," Fuzzy-IEEE'02, Vol.1, pp.488-493, 2002

[4] Ken Nozali, Hisao Ishibuchi, Hideo Tanaka, "Adaptive Fuzzy Rule-Based Classification Systems," IEEE Transations on Fuzzy Systems, Vol.4, No.3, pp.238-250, 1996

[5] J. Roubos, M. Setnes, J. Abonyi,, "Learning Fuzzy Classification Rules from Labeled Data," International Journal of Information Sciences, 150(1-2), pp.77-93, 2003

[6] J. Abonyi, J. Roubos, F. Szeifert, "Data-driven generation of compact, accurate, and linguistically sound fuzzy classifiers based on a decision-tree initialization," International Journal of Approximate Reasoning, 31(1), pp.1-21, 2003

[7] Janikow, C. Z., Fajfer, M., "Fuzzy partitioning with FID3.1," 18th International Conference of the North American, NAFIPS, pp.467-471, 1999

[8] 김성은, 류정우, 김명원,"효율적인 지도 퍼지 군집화를 위한 휴리스틱 분할 진화알고리즘," 한국정보과학회 춘계학술대회 논문집, 제32권, 제1호(B), pp.667-669, 2005

[9] Myung Won Kim, Joung Woo Ryu, "Optimized Fuzzy Classification using Genetic Algorithm,"LNAI 3613, pp. 392-401, 2005

[10] X.-Z. Wang, D. S. Yeung, E. C. C. Tsang, "A Comparative Study on Heuristic Algorithms for Generating Fuzzy Decision Trees," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 31, NO. 2, pp. 215-226, 2001